CLUSTERING BY OPTIMUM PATH FOREST AND ITS APPLICATION TO AUTOMATIC GM/WM CLASSIFICATION IN MR-T1 IMAGES OF THE BRAIN

Fábio A. M. Cappabianco
Alexandre X. Falcão

State University of Campinas
Institute of Computing
Av. Albert Einstein, 1251,
CEP 13084-851, Campinas, SP, Brasil

Leonardo M. Rocha

State University of Campinas
School of Electrical and Computing Engineering
Av. Albert Einstein, 400,
CEP 13083-970, Campinas, SP, Brasil

ABSTRACT

A new approach to identify clusters as trees of an optimum-path forest has been presented. We are extending the method for large datasets with application to automatic GM/WM classification in MR-T1 images of the brain. The method is computed for a few randomly selected voxels, such that GM and WM define two optimum-path trees. The remaining voxels are classified incrementally, by identifying which tree would contain each voxel if it were part of the forest. Our method produces accurate results on phantom and real images, similarly to those obtained by the state-of-the-art, does not rely on templates, and takes less than 1.5 minute on modern PCs.

Index Terms— Medical image processing, image foresting transform, improved mean-shift algorithm, graph-cut measures, MR image segmentation.

1. INTRODUCTION

We have presented an approach for data clustering based on optimum-path forest [1]. The samples are nodes of a graph, whose arcs connect k-nearest neighbors in the feature space. The graph is weighted on the nodes by density values, forming a discrete probability density function (pdf), which is computed from the distances (arc weights) between the feature vectors of the adjacent samples. The best k is found by minimizing a graph-cut measure and the maximization of a *path-value function* outputs an *optimum-path forest* (OPF), where each tree (cluster) is rooted at a maximum of the pdf. The method is more general and improves the mean-shift algorithm [2] for data clustering in robustness and number of irrelevant clusters. It also extends the image foresting transform [3] from the image domain to the feature space.

Graph-based clustering methods for image analysis usually define the arcs between pixels within a small adjacency radius [1, 2, 4, 5], due to the large number of samples (even for 2D images). The present work extends the OPF clustering to large datasets with no adjacency constraints in the image domain. The impact of this result becomes evident when we apply the method for tissue classification in MR-T1 images of the brain. It takes less than 1.5 minute on modern PCs with no need for templates and no user intervention. For a volume containing only gray matter (GM) and white matter (WM) voxels, a subset of voxels (typically less than 500 samples) is randomly selected in the brain to form the k-nn graph and an optimum-path forest is computed with only two trees, each representing a cluster of GM or WM. The remaining voxels are classified incrementally, by identifying which tree would contain each voxel if it were part of the forest.

The brain (GM + WM) can be isolated in MR-T1 images either interactively [6] or automatically [7], within a few seconds, and with no need of templates and adjustment of parameters. In these approaches, the CSF is eliminated by removing voxels below the Otsu's threshold computed on the original image. Other approaches exist for MR-image segmentation of the brain, including GM and WM separation. Despite of the known criticism [8, 9], most of them relies on templates to obtain at least prior information [10–12]. Some of them take several minutes to separate GM and WM [13, 14] and others require multiple imaging modalities [15, 16].

The proposed approach can obtain similar accuracies from a single modality. The next sections present a review on the OPF clustering, its extension to large datasets, its application to GM/WM classification, the experiments with phantoms and real images, results and conclusions.

2. OPTIMUM PATH FOREST CLUSTERING

Let $\mathcal{N}$ be a dataset such that for every sample $s \in \mathcal{N}$ there is a feature vector $\vec{v}(s)$. Let $d(s,t)$ be the distance between s and t in the feature space (e.g., $d(s,t) = \|\vec{v}(t) - \vec{v}(s)\|$). The fundamental problem in data clustering is to identify natural groups in $\mathcal{N}$. In GM/WM classification, there are two groups (possibly with overlap) and a distinct label must be assigned to the samples of each group.

The authors thank FAPESP and CNPq

A graph $(\mathcal{N}, \mathcal{A})$ is defined such that the arcs $(s,t) \in \mathcal{A}$ connect k-nearest neighbors in the feature space. The arcs are weighted by $d(s,t)$ and the nodes $s \in \mathcal{N}$ are weighted by a density value $\rho(s)$.

$$\rho(s) = \frac{1}{\sqrt{2\pi\sigma^2}|\mathcal{A}(s)|} \sum_{\forall t \in \mathcal{A}(s)} \exp\left(\frac{-d^2(s,t)}{2\sigma^2}\right) \quad (1)$$

where $|\mathcal{A}(s)| = k$, $\sigma = \frac{d_f}{3}$, and d_f is the maximum arc weight in $(\mathcal{N}, \mathcal{A})$. This parameter choice considers all nodes for density computation, since a Gaussian function covers most samples within $d(s,t) \in [0, 3\sigma]$. The traditional method to estimate a probability density function (pdf) is by Parzen-window. Equation 1 can provide a Parzen-window estimation based on isotropic Gaussian kernel when we define the arcs by $(s,t) \in \mathcal{A}$ if $d(s,t) \leq d_f$. This choice, however, presents problems with the differences in scale and sample concentration. Solutions for this problem lead to adaptive choices of d_f depending on the region of the feature space [17]. By taking into account the k-nearest neighbors, we are handling different concentrations and reducing the scale problem to the one of finding the best value of k within $[1, k_{\max}]$, for $1 \geq k_{\max} \leq |\mathcal{N}|$. Our solution considers the minimum graph cut provided by the clustering results for $k \in [1, k_{\max}]$, according to a measure $C(k)$ suggested by Shi and Malik [4].

$$C(k) = \sum_{i=1}^{c} \frac{W_i'}{W_i + W_i'}, \quad (2)$$

$$W_i = \sum_{\forall (s,t) \in \mathcal{A} | L(s)=L(t)=i} \frac{1}{d(s,t)}, \quad (3)$$

$$W_i' = \sum_{\forall (s,t) \in \mathcal{A} | L(s)=i, L(t)\neq i} \frac{1}{d(s,t)}, \quad (4)$$

where $L(t)$ is the label of sample t, W_i' uses all arc weights between cluster i and other clusters, and W_i uses all arc weights within cluster $i = 1, 2, \ldots, c$. Figure 1a shows an example with $|\mathcal{N}| = 340$ samples, which form a few clusters with different sample concentrations in a 2D feature space. Depending on the scale, there are one, three, four, or five natural groups. If $k_{\max} \geq 150$, then the minimum cut will occur when all samples are grouped into a single cluster. The minimum cut for $k_{\max} = 100$ identifies four clusters with a best $k = 37$ (Figure 1b), and by limiting the search to $k_{\max} = 30$, the minimum cut identifies five clusters with a best $k = 29$ (Figure 1c). The clustering results are obtained by extending the image foresting transform [3] (IFT) from the image domain to the feature space, as follows. A path π_t is a sequence of adjacent samples starting from a root $R(t)$ and ending at a sample t, being $\pi_t = \langle t \rangle$ a trivial path and $\pi_s \cdot \langle s,t \rangle$ the concatenation of π_s and arc (s,t). Among all possible paths π_t with roots on the maxima of the pdf, we wish to find a path whose the lowest density value along it is maximum. Each maximum should then define an influence zone (cluster) by selecting the samples that are more strongly connected to it, according to this definition, than to any other maximum. More formally, we wish to maximize $f(\pi_t)$ for all $t \in \mathcal{N}$ where

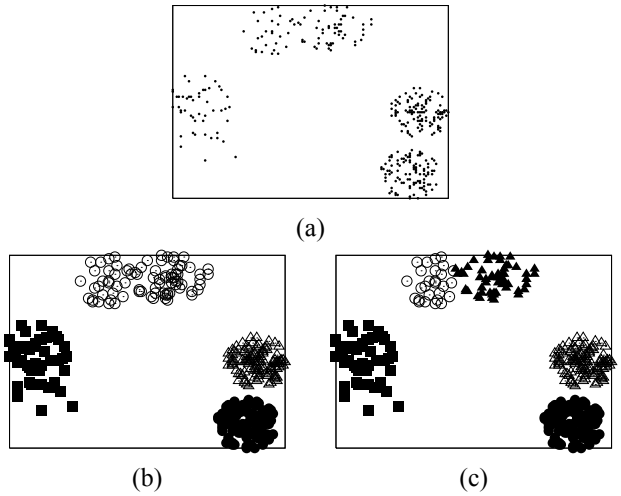


Fig. 1. (a) Feature space with distinct sample concentration per cluster. We can identify different cluster numbers depending on the scale. Reasonable solutions are (b) four and (c) five clusters, where two touch each other.

$$f(\langle t \rangle) = \begin{cases} \rho(t) & \text{if } t \in \mathcal{R} \\ \rho(t) - \delta & \text{otherwise} \end{cases}$$

$$f(\langle \pi_s \cdot \langle s,t \rangle \rangle) = \min\{f(\pi_s), \rho(t)\} \quad (5)$$

for $\delta = \min_{\forall (s,t) \in \mathcal{A} | \rho(t) \neq \rho(s)} |\rho(t) - \rho(s)|$ and $\mathcal{R}$ being a root set with one element for each maximum of the pdf. Higher values of delta reduce the number of maxima. We are setting $\delta = 1.0$ and scaling real numbers $\rho(t) \in [1, 1000]$ in this work. The IFT algorithm maximizes $f(\pi_t)$ such that the optimum paths form an optimum-path forest — a predecessor map P with no cycles that assigns to each sample $t \notin \mathcal{R}$ its predecessor $P(t)$ in the optimum path from $\mathcal{R}$ or a marker nil when $t \in \mathcal{R}$.

Algorithm 1 – CLUSTERING BY OPTIMUM PATH FOREST

INPUT: Graph $(\mathcal{N}, \mathcal{A})$ and function ρ.
OUTPUT: Label map L, path-value map V, forest P.
AUXILIARY: Priority queue Q, variables tmp and $l \leftarrow 1$.

For all $s \in \mathcal{N}$, set $P(s) \leftarrow nil$, $V(s) \leftarrow \rho(s) - \delta$, insert s in Q.
While Q is not empty, do
  Remove from Q a sample s such that $V(s)$ is maximum.
  If $P(s) = nil$, then
    Set $L(s) \leftarrow l$, $l \leftarrow l + 1$, and $V(s) \leftarrow \rho(s)$.
  For each $t \in \mathcal{A}(s)$ and $V(t) < V(s)$, do
    Compute $tmp \leftarrow \min\{V(s), \rho(t)\}$.
    If $tmp > V(t)$ then
      Set $L(t) \leftarrow L(s)$, $P(t) \leftarrow s$, $V(t) \leftarrow tmp$.
      Update position of t in Q.

Agorithm 1 identifies one root in each maximum of the pdf ($P(s) = nil$ in Line 4 implies $s \in \mathcal{R}$), assigns to each root a distinct label in Line 5, and computes the influence zone (cluster) of each root as an optimum-path tree in P, such that the nodes of the tree receive the same label of its root in a map L (Line 9). It also outputs the optimum path-value map V and forest P. It is more robust than the mean-shift [2] because it does not depend on pdf gradients, uses a k-nn graph and assigns a single label per maximum, even when the maximum is a connected component in $(\mathcal{N}, \mathcal{A})$. It is more general because the choice of $f(\langle t \rangle)$ can reduce irrelevant maxima (clusters).

3. EXTENSION TO LARGE DATASETS

Algorithm 1 takes $O(k|\mathcal{N}| + |\mathcal{N}| \log \mathcal{N})$, when Q is a binary heap, and the estimation of the best k requires its computation several times. This can take several minutes on modern PCs for $|\mathcal{N}| > 1000$. The problem becomes unsurmountable for 2D/3D images. In [1], the number of arcs are considerably reduced by defining $\mathcal{A}$ as $(s,t) \in \mathcal{A}$ if $\|t - s\| \leq h_i$ and $d(s,t) \leq h_f$ for image segmentation. The result, however, becomes a compromise between the choice of h_i, whose smaller values increase the number of irrelevant clusters, and the choice of $f(\langle t \rangle)$, which can reduce this number. We are avoiding any constraint in the image domain, by incremental implementation of the method on k-nn graphs.

The extension is based on a random selection of a set $\mathcal{N}' \subset \mathcal{N}$. Let V and L be the optimum maps from Algorithm 1 computed on the best k-nn graph $(\mathcal{N}', A)$. A sample $t \in \mathcal{N} \backslash \mathcal{N}'$ can be classified in one of the clusters by identifying which root would offer it an optimum path as though it were part of the forest. By considering the k-nearest neighbors of t in $\mathcal{N}'$, we can use Equation 1 to compute $\rho(t)$, evaluate the optimum paths $\pi_s \cdot \langle s,t \rangle$, and select the one that satisfies

$$V(t) = \max_{\forall (s,t) \in \mathcal{A}} \{\min\{V(s), \rho(t)\}\} \quad (6)$$

Let the node $s^* \in \mathcal{N}'$ be the one that satisfies Equation 6. The classification simply assigns $L(s^*)$ as the cluster of t.

4. APPLICATION TO GM/WM CLASSIFICATION

An MR-T1 image of the brain is a pair $(\mathcal{N}, I)$, where $\mathcal{N}$ is the voxel set and $I(t)$ is the voxel intensity. The problem consists of finding two clusters ($c = 2$), GM and WM. The subgraph $(\mathcal{N}', \mathcal{A})$ is created by sampling 0.02% of the voxels from $\mathcal{N}$, such that 0.01% of these voxels have values below the mean brightness inside the brain and 0.01% have values above it. This allows a fair amount of samples from both GM and WM. The feature vector $\vec{v}(t)$ consists of the value $I(t)$ and the values of its six closest neighbors in the image domain. When the neighbor is out of the brain, we repeat $I(t)$ in the vector. The best value of k is found within $[1, k_{\max}]$. We set $k_{\max} = 50$ due to the scale problem created by inhomogeneity and partial volume. The method usually finds two clusters in this range, but it is possible to appear more than two. In such a case, we force two clusters by assigning a GM label to those with mean brightness below the mean intensity in the brain and a WM label otherwise. Equation 6 is then evaluated to classify the remaining voxels $t \in \mathcal{N} \backslash \mathcal{N}'$.

5. EVALUATION

We selected 8 phantom images with $181 \times 217 \times 181$ voxels from the Brainweb database [1], with noise from 3%, 5%, 7%, and 9%, and inhomogeneity 20% and 40%, respectively. The ground-truth image is available and the similarity between each result and ground-truth was computed using the Dice metric. We executed the method 10 times for each phantom using different randomly selected sets $\mathcal{N}'$ and computed the mean and standard deviation of the Dice similarities. We have also performed the same experiment for the first 8 real images (with 9-bit intensity values) from the IBSR dataset [2].

Table 1 presents the Dice similarities for the phantom images. These results are good and equivalent to those obtained by recent approaches [14]. In the case of real images (Table 2), our method obtained similarities 0.90 for GM and 0.86 for WM on average, against 0.80 for GM and 0.88 for WM reported in [14]. This difference in favor of our approach for GM classification is a relevant result for medical studies.

We have also measured the computational time for unsupervised learning (OPF clustering including the best k estimation) and classification. The longest execution time was 84 seconds on a 3GHz Pentium IV PC. Given that the method proposed in [14] used a 2.8GHz Pentium IV PC, our method is about 30 times faster than that approach.

Phantom Dataset	Dice simil. for GM		Dice simil. for WM	
	mean	std. dev.	mean	std. dev.
1(3%,20%)	95.15%	0.17	93.43%	0.19
2(5%,20%)	95.10%	0.17	93.40%	0.20
3(7%,20%)	94.36%	1.03	92.55%	0.93
4(9%,20%)	94.06%	0.27	91.93%	0.54
5(3%,40%)	90.90%	1.28	88.30%	0.64
6(5%,40%)	91.23%	1.25	88.19%	0.67
7(7%,40%)	91.10%	0.72	87.77%	0.81
8(9%,40%)	90.66%	1.21	87.03%	0.73

Table 1. Results with phantoms: mean and standard deviation of the Dice similarities for GM and WM.

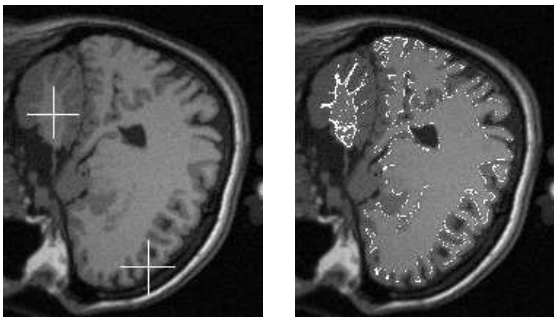
The inhomogeneity seems to be the greatest challenge. It is not difficult to find different regions in the brain where GM and WM have similar intensities (Figure 2a). The errors were concentrated on the boundary between GM and WM, being

[1] URL: http://www.bic.mni.mcgill.ca/brainweb
[2] URL: www.cma.mgh.harvard.edu/ibsr

IBSR Dataset	Dice simil. for GM		Dice simil. for WM	
	mean	std. dev.	mean	std. dev.
1	92.22%	0.87	84.98%	2.03
2	90.99%	2.93	86.55%	2.93
3	93.86%	0.14	86.07%	0.85
4	88.19%	5.97	85.99%	3.31
5	90.20%	1.73	84.59%	1.40
6	85.02%	4.21	83.00%	3.32
7	91.22%	3.35	87.39%	2.79
8	88.46%	4.39	86.05%	3.41

Table 2. Results with real images: mean and standard deviation of the Dice similarities for GM and WM.

clearly more sensitive to the inhomogeneity variation than to the noise variation (Figure 2b).



(a) (b)

Fig. 2. (a) The markers show WM and GM voxels with the same intensity 1642. (b) The errors (white overlay) are concentrated on the boundary between GM and WM.

It also is interesting to point that our method did not improve accuracy significantlty by increasing the sampling rate from 0.02% to 0.12%. Further improvements should be done by choice of better feature vectors and inhomogeneity correction.

6. CONCLUSION

A very efficient extension of the OPF clustering algorithm [1] was proposed to large datasets and validated for GM/WM classification. The experiments involved phantom and real images of the brain. The classification results were good and similar to those reported by recent approaches [14]. However, the proposed method is about 30 times faster, do not rely on templates, and is significantly more accurate for GM classification in real images.

Our future goals will be to improve feature selection and to evaluate possible variants of the method, which may increase accuracy with minimum user intervention.

7. REFERENCES

[1] L.M. Rocha, A.X. Falcão, and L.G.P. Meloni, "A robust extension of the mean shift algorithm using optimum path forest," in *Proc. of the 12th Intl. Workshop on Combinatorial Image Analysis*, 2008, to appear.

[2] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE TPAMI*, vol. 17, no. 8, pp. 790–799, Aug 1995.

[3] A. X. Falcão, J. Stolfi, and R. A. Lotufo, "The image foresting transform: theory, algorithms, and applications," *IEEE TPAMI*, vol. 26, no. 1, pp. 19–29, Jan 2004.

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, Aug 2000.

[5] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. of the ICCV*, 2001, vol. 1, pp. 105–112.

[6] A. X. Falcão and F. P. G. Bergo, "Interactive volume segmentation with differential image foresting transforms," *IEEE TMI*, vol. 23, no. 9, pp. 1100–1108, Sep 2004.

[7] F.P.G. Bergo, A.X. Falcão, P.A.V. Miranda, and L.M. Rocha, "Automatic image segmentation by tree pruning," *Journal of Mathematical Imaging and Vision*, 2007, to appear.

[8] F.L. Bookstein, "Voxel-based morphometry should not be used with imperfectly registered images," *Neuroimage*, vol. 14, no. 6, pp. 1454–1462, 2001.

[9] N. Thacker, "A critical analysis of voxel-based morphometry," Tech. Rep., Tina Memo, 2005.

[10] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden markov randomfield model and the expectation-maximization algorithm," *IEEE TMI*, vol. 20, no. 1, pp. 45–57, 2001.

[11] V. Grau, A.U.J. Mewes, M. Alcaniz, R. Kikinis, and S.K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE TMI*, vol. 23, no. 4, pp. 447–458, 2004.

[12] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "A unifying framework for partial volume segmentation of brain mr images," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 1, pp. 105–119, 2003.

[13] S. Bricq, C. Collet, and J.P. Armspach, "Triplet markov chain for 3D MRI brain segmentation using a probabilistic atlas," *3rd IEEE International Symposium on Biomedical Imaging: From Macro to Nano, 2006*, pp. 386–389, 2006.

[14] SP. Awate, T. Tasdizen, N. Foster, and RT. Whitaker, "Adaptive markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification," *Medical Image Analysis*, vol. 10, no. 5, pp. 726–39, 2006.

[15] M. Prastawa, J.H. Gilmore, W. Lin, and G. Gerig, "Automatic segmentation of MR images of the developing newborn brain," *Medical Image Analysis*, vol. 9, no. 5, pp. 457–66, 2005.

[16] K.M. Pohl, S. Bouix, M. Nakamura, T. Rohlfing, R.W. McCarley, R. Kikinis, W.E.L. Grimson, M.E. Shenton, and W.M. Wells, "A hierarchical algorithm for MR brain image parcellation," *IEEE TMI*, vol. 26, no. 9, pp. 1201–1212, 2007.

[17] D. Comaniciu, "An algorithm for data-driven bandwidth selection," *IEEE TPAMI*, vol. 25, no. 2, pp. 281–288, 2003.